Discriminating between disease-causing and neutral non-frameshifting micro-INDELs by support vector machines by means of integrated sequence- and structure-based features

Huiying Zhao[1], Yuedong Yang[1,2], Hai Lin[2], Xinjun Zhang[2], Matthew Mort[4], David N. Cooper[4], Yunlong Liu[2,3,*], and Yaoqi Zhou[1,2,*]

[1]School of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202, USA

[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

[3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA

[4]Institute of Medical Genetics, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

ABSTRACT

Micro-INDELs (insertions or deletions of ≤20 bp) constitute the second most frequent class of human gene mutation after single nucleotide variants. A significant portion of exonic INDELs are non-frameshifting (NFS), serving to insert or delete a discrete number of amino-acid residues. Despite the relative abundance of NFS-INDELs, their damaging effect on protein structure and function has gone largely unstudied whilst bioinformatics tools for discriminating between disease-causing and neutral NFS-INDELs remain to be developed. We have developed such a technique (*DDIG-in*; Detecting DIsease-causing Genetic variations due to INDELs) by comparing the properties of disease-causing NFS-INDELs from the Human Gene Mutation Database (HGMD) with putatively neutral NFS-INDELs from the 1,000 Genomes Project. Having considered 58 different sequence- and structure-based features, we found that predicted disordered regions around the NFS-INDEL region had the highest discriminative capability (disease versus neutral) with an Area Under the receiver-operating characteristic Curve (AUC) of 0.82 and a Matthews Correlation Coefficient (MCC) of 0.56. All features studied were combined by support vector machines (SVM) and selected by a greedy algorithm. The resulting SVM models were trained and tested by ten-fold cross-validation on the microdeletion dataset and independently tested on the microinsertion dataset and *vice versa*. The final SVM model for determining NFS-INDEL disease-causing probability was built on non-redundant datasets with a protein sequence identity cutoff of 35% and yielded an MCC value of 0.68, an accuracy of 84% and an AUC of 0.89. Predicted disease-causing probabilities exhibited a strong negative correlation with the average minor allele frequency (correlation coefficient, -0.84). *DDIG-in*, available at http://sparks.informatics.iupui.edu, can be used to estimate the disease-causing probability for a given NFS-INDEL.

Mentor: Yaoqi Zhou and Yunlong Liu.